# Predicting Skill Qualification Test Item Difficulty from Judgments

Douglas Macpherson
Research Psychologist
US Army Research Institute
for the Behavioral and Social Sciences

Judgments of item difficulty by small groups of three to six non-commissioned officers were compared with observed item difficulties among soldiers in three military occupational specialties representing infantry, engineer and administrative career fields. Linear correlations between average judgments and observed difficulties were on the order of .50, but the scatter plots were triangular in appearance because objectively easy items were rarely judged to be difficult while objectively difficult items yielded a wide range of judged difficulties. Hence sets of items showing wide and fairly flat distributions of difficulty had been judged to be skewed toward the easy end of the difficulty distribution. These analytic observations suggest that NCOs involved in test construction may be making tests more difficult than they believe and that NCOs as trainers preparing soldiers for their SQTs may be underestimating the need for training. If the triangular relationship between judged and observed difficulty is confirmed in larger samples of items, then a simple expectancy table method might be used to predict objective test difficulty and training need.

# Predicting Skill Qualification Test
# Item Difficulty from Judgments

## INTRODUCTION

Teams of experienced soldiers, oriented in criterion referenced test development workshops and assisted by civilian test psychologists, produce Skill Qualification Tests (SQT). The procedures and policy guidelines were recently reviewed by a well known authority on criterion referenced testing (Hambleton, 1981). His assessment was that the prescribed procedures were excellent. His only reservations concerned the lack of reliability information and the obvious possibility of failure in execution. However, Hambleton apparently accepted that standards can be set by testing proficient and non-proficient personnel on each task and then selecting cut points which discriminate between them without considering the total scores.

This study examined one possible form of test development bias. We wanted to determine if supervisors exhibited "item leniency," and judged the items to be easier for enlisted personnel than the items were found to be. Thus we asked supervisors of a combat MOS (11H, TOE Gunner), an engineering MOS (12C, Bridge Construction Crewman), and an administrative MOS (71L, Clerk) to estimate what percentage of their troops would pass each item of the Skill Component (SC) portion of the appropriate SQT.

## METHOD

Subjects: The subjects were 10E5 - E7 NCO supervisors and 2 E4 Acting Supervisors for MOS 11H, 12C, and 71L at Forts Bragg, Carson and Hood. They were distributed as follows: (a) six 11H, (b) three 12C, and (c) three 71L. All supervisors held the appropriate MOS or a closely related MOS as their primary or secondary MOS. In addition they had held the MOS for an average of 4 years (range 10 months - 15 years) and had supervised for an average of 4 1/2 years (range 8 months - 18 years).

Procedures: Supervisors at the three posts reviewed the SC which had been administered to their subordinates within the last year. They were instructed to judge what percentage of their soldiers could get each item correct and to write the estimate in the SQT. Because of the length of the tests each supervisor was asked to rate a specified set of subtests which usually comprised about one half of the test. Some supervisors, however, completed all the items. The rating task required about a half hour to complete.

Criterion data: Item analyses for each SQT were provided by SQT Management Directorate of TRADOC. These analyses provided the total number in of testees in each sample and the percentage of testees selecting each alternative for each item.

## RESULTS

We determined the reliability of the criterion data for one SQT for which we had two samples. The results for each SQT were examined for rating reliability and rating validity. The rating reliability for each SQT is presented in terms of the total number of raters used. The reliabilities are then recalculated to the common metric of the correlation to be expected between a pair of raters by using the Spearman Brown Prophesy formula. Similarly rating validities are presented as the correlation of the mean rater estimates with the criterion, and then corrected for the attenuation due to predictor unreliability. As a result of these adjustments, reliabilities and validities may be compared across the three SQT. Finally, the validity results for the three SQT were combined. The analyses are presented in the form of two way tables.

Criterion Reliability. The 11H SQT results were obtained as two samples of 789 and 922 soldiers respectively. These samples consisted of all 11H2180 scores which had been transcribed by SMD before 21 Dec 80 and between 21 Dec and 22 Mar 81. The product moment correlation for the proportions of soldiers in the two groups who passed the items was .99. Table 1 displays the high correlation obtained and demonstrates that the test contained a relatively uniform spread of difficulty levels. Table 1 also indicated that there was little evidence for change in scores over the course of the two administrations. This was supported by the regression equation: Second Sample = .05 + .924 First Sample. The mean and sigma for the two samples combined were 61 and 20 respectively. Multiple samples were not available for the other SQT therefore criterion reliability analyses were not possible for those SQT.

**Table 1. Reliability of Item Analysis Difficulty Levels for 61 Item Test 11H2180 SC**

| | | First Sample | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | 10<br>19 | 20<br>29 | 30<br>39 | 40<br>49 | 50<br>59 | 60<br>69 | 70<br>79 | 80<br>89 | 90<br>100 | n | % |
| Second Sample | 90-99 | | | | | | | | 1 | 4 | 5 | 8 |
| | 80-89 | | | | | | | 1 | 9 | | 10 | 16 |
| | 70-79 | | | | | | 2 | 6 | 2 | | 10 | 16 |
| | 60-69 | | | | | 1 | 3 | 3 | | | 7 | 11 |
| | 50-59 | | | | 1 | 8 | | | | | 9 | 15 |
| | 40-49 | | | 5 | 4 | 1 | | | | | 10 | 16 |
| | 30-39 | | 2 | 4 | 1 | | | | | | 7 | 11 |
| | 20-29 | 2 | 1 | | | | | | | | 3 | 5 |
| | 10-19 | | | | | | | | | | 0 | 0 |
| | n | 2 | 3 | 9 | 6 | 10 | 5 | 10 | 12 | 4 | 61 | |
| | % | 3 | 5 | 15 | 10 | 16 | 8 | 16 | 20 | 7 | | 100 |

11H2180 Reliability and Validity. Six supervisors of 11H troops rated all subtests. These raters were divided into two groups and the mean difficulty ratings for the two groups were compared. The split half correlation was .78 (expected pairwise correlations of .54). Table 2 confirms that the two groups generated comparable estimates.

Table 2. Split Half Reliability of Estimated Item Difficulty Levels for SQT 11H2180 (6 raters)

| | | Group A (3 raters) | | | | |
|---|---|---|---|---|---|---|
| | %<br>Correct | 20-39 | 40-59 | 60-79 | 80-99 | n<br>% |
| Group B (3 raters) | 80-99 | | 3 | 3 | 19 | 25<br>42 |
| | 60-70 | 1 | 2 | 5 | 6 | 14<br>23 |
| | 40-59 | | 1 | 6 | | 7<br>12 |
| | 20-39 | 6 | 8 | | | 14<br>23 |
| | n<br>% | 7<br>12 | 14<br>23 | 14<br>23 | 25<br>42 | 60<br>100 |

The validity correlation for the mean item difficulty estimates was .49. When corrected for predictor attenuation the r became .55. Although the standard deviations were about equal, 23% vs 21%, the mean estimated percent correct was 68% vs 61% for the actual test. Thus the NCOs tended to agree with each other but vary from the criterion and underestimate the difficulty of the test items. Furthermore, the underestimates were not uniform. The supervisors correctly estimated the difficulty of the easy items, but not the difficult items. As a result the upper left quadrant was empty, whereas the lower right quadrant contained 1/5 of the observations and about the same density as the validity diagonal. Since the quadrant boundaries represent approximately a 1.5 SD difference from valid predictions, the significance of the effect is apparent. These effects appear in Table 3.

Table 3. Validity of Item Difficulty Estimates for Test 11H2180

| | | Estimated % Correct (6 raters) | | | | |
|---|---|---|---|---|---|---|
| | % | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| Observed % Correct (n = 1711) | 80-100 | | | 5 | 12 | 17<br>28 |
| | 60-79 | | 1 | 6 | 6 | 13<br>22 |
| | 40-59 | 7 | 5 | 2 | 4 | 18<br>30 |
| | 20-39 | 3 | 1 | 6 | 2 | 12<br>20 |
| | n<br>% | 10<br>17 | 7<br>12 | 19<br>32 | 24<br>40 | 60<br>100 |

12C2180 Reliability and Validity. Three NCOs for 12C soldiers rated all subtests. The consistency of these ratings was determined to be .40 with Cronbach's Alpha; the pairwise correlation was estimated to be .31. The correlation across 121 items of these NCOs with the observed performance of soldiers on SQT 12C2180 SC was .50 (.79 corrected for predictor attenuation). The SDs for the NCO and criterion data were both 20%, whereas the means were 69% and 64% for the raters and the criterion respectively. The dispersion of the ratings is displayed in Table 4. The row marginal totals indicated that the test item difficulties were uniformly distributed.

Table 4. Validity of Item Difficulty Estimates for Test 12C2180 (SC)

| | | Estimated % Correct (3 raters) | | | | |
|---|---|---|---|---|---|---|
| | % | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| Observed % Correct | 80-100 | 1 | 3 | 9 | 18 | 31<br>26 |
| | 60-79 | 2 | 7 | 10 | 13 | 32<br>26 |
| | 40-59 | 2 | 11 | 9 | 13 | 35<br>29 |
| | 20-39 | 5 | 8 | 7 | 1 | 21<br>17 |
| | 0-19 | 2 | | | | 2<br>2 |
| | n<br>% | 12<br>10 | 29<br>24 | 35<br>29 | 45<br>37 | 121<br>100 |

In contrast, the column marginal totals indicated that the raters tended to generate a negatively skewed distribution. Again there tended to be many fewer responses in the upper left than in the lower right quadrant.

71L2180 Reliability and Validity. The 71L2180 Skill Component introduced additional complexities. It utilized multiple correct responses to some items. Since the multiple correct answer format is no longer used in SQT construction we excluded these items from the analyses.

Only three raters were available for this SQT. Rater #3 completed the entire questionnaire whereas two other raters completed the first and second halves of the SQT respectively.

The reliability of the three raters was estimated by correlating the estimates of rater #3 with the combination of the other raters separately. The reliability correlation was found to be .48. We regarded this as a pairwise correlation and did not reduce it further.

The validity was estimated by correlating the means of the appropriate pairs of ratings with the criterion scores. The mean and SD for the difficulty estimates were 65% and 21% respectively. However, the test was actually more difficult than the others examined. Thus the criterion mean was 50%. However the criterion variance was a typical 19%.

The validity correlation for the three raters was a low .28. Adjusting this correlation for rater unreliability increased the validity correlation to .40.

The distribution of the items is presented in Table 5. Table 5 displays the usual pattern of underestimating the item difficulties. The raters tended to rate the easy items as easy. However, they failed to demonstrate such consistency in their ratings of the difficult items. Instead they tended to use all rating values in evaluating the difficult items. Again, if they identified an item as difficult it was difficult.

**Table 5. Validity of Item Difficulty Estimates For Test 71L2180 (SC) Single Correct Response Items**

| | | Estimated % Correct (mean of 2 raters) | | | | | |
|---|---|---|---|---|---|---|---|
| | % | 0-19 | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| Observed % Correct | 80-100 | | | | 1 | 2 | 3<br>4 |
| | 60-79 | 1 | 1 | 2 | 4 | 9 | 17<br>25 |
| | 40-59 | | 5 | 6 | 13 | 7 | 31<br>46 |
| | 20-39 | | 2 | 3 | 4 | 3 | 12<br>17 |
| | 0-19 | | 1 | 2 | 1 | 1 | 5 |
| | n<br>% | 1<br>1 | 9<br>13 | 13<br>19 | 23<br>34 | 22<br>32 | 68<br>100 |

Item Difficulty Estimates Summarized. The results for the three SQT are summarized at the item level in Table 6. The mean and standard deviation for the observed difficulty across the three SQT were 58% and 22% respectively. Similarly, the mean and SD for the difficulty estimates were 68% and 21% respectively. Table 6 is in expectancy table format. The number in each cell is the percent of items in the table that are in the cell.

Table 6. Validity of Item Difficulty Estimates for Test 71L2180 (SC) Multiple Correct Response Items

| | Estimated % Correct | | | | | | |
|---|---|---|---|---|---|---|---|
| | % | 0-19 | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| Observed % Correct | 80-100 | | * | | 6 | 13 | 20 |
| | 60-79 | * | 1 | 4 | 8 | 11 | 25 |
| | 40-59 | | 6 | 9 | 10 | 10 | 34 |
| | 20-39 | | 4 | 5 | 7 | 2 | 18 |
| | 0-19 | | 1 | 1 | | * | 2 |
| | % | * | 12 | 20 | 31 | 37 | 100 |

* Only one response

The marginal totals demonstrated that the item difficulties were relatively uniformly distributed through the range 20% - 100%. However, the negative skew of the estimates is clearly seen. Thus, the pattern of errors was non-random and non-linear. As expected, Table 6 exhibits a nearly triangular matrix. The upper left quadrant contains only two percent of the sample whereas the lower right quadrant contains 19% of the sample and the cell densities resemble the cell densities on the main diagonal. Across the three MOS, easy items were rated as easy whereas the difficult items received almost any rating. However, if a supervisor rated an item as difficult then the item was difficult

## DISCUSSION

This preliminary research demonstrated that NCOs can predict the item by item test performance of their troops to a moderate degree. However, they do tend to underestimate the difficulty of more than 40% of the items. If our conjecture that the test developers would respond like NCOs is correct, then this study does suggest that one reason for low SQT scores is that the tests are harder than intended from a normative point of view. Furthermore this is not detected during SQT pretesting for good reasons. The SOP describes a procedure in which each subtest is tested independently of the other subtests using personnel who are expert and non-expert on that subtest. As a result no estimates of test total scores are available prior to field use of the test.

There is a second way in which NCO underestimation of item difficulty can yield low SQT scores. The NCOs who performed the ratings were responsible for training troops to take the test. They were provided with the topics to be covered in the test and sample items for each topic. They may have trained their troops to the apparent difficulty level of the test. As a result they may tend to provide insufficient training for the SC portion of the test.

The combination of the two effects - tests more difficult than intended, and inadequate training for the test - could account for the disappointing scores observed on many SQT.

## REFERENCE NOTE

Hambleton, R. D. Psychometric methods for the Skill Qualification Test - Final Report. Amherst, Mass.: University of Massachusetts, 1981.

## REFERENCES

Department of the Army. Guidelines for Development of Skill Qualification Tests (SQT) Policy and Procedures. TRADOC Pam 351-2 (Draft) Ft Monroe, VA: Headquarters, US Army Training and Doctrine Command, Updated (Supercedes TRADOC Pam 351-2, 1 Dec 1977).

Department of the Army. Skill Qualification Test (SQT) Policy and Procedures. TRADOC Reg 351-2. Ft. Monroe, VA: Headquarters, US Army Training and Doctrine Command, 27 April 1980.

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. J. Education Measurement, 1978, 15, 277-290.